Computer Science Colloquium

Resource-Efficient Execution for Deep Learning

Speaker: Deepak Narayanan

Host: Abhishek Bhattacharjee

Tuesday, February 16, 2021
4:00 p.m.

Zoom Presentation

ABSTRACT:

Deep Learning models have enabled state-of-the-art results across a broad range of applications; however, training these models is extremely time- and resource-intensive, taking weeks on clusters with thousands of expensive accelerators in the extreme case. In this talk, I will describe two systems that improve the resource efficiency of model training. The first system, PipeDream, proposes the use of pipelining to accelerate distributed training. Pipeline parallelism facilitates model training with lower communication overhead than previous methods while still ensuring high compute resource utilization. Pipeline parallelism also enables the efficient training of large models that do not fit on a single worker. Pipeline parallelism is being used at Facebook, Microsoft, OpenAI, and Nvidia for efficient large-scale model training. The second system, Gavel, determines how resources in a shared cluster with heterogeneous compute resources (e.g., different types of hardware accelerators) should be partitioned among different users to optimize objectives specified over multiple training jobs. Gavel can improve various scheduling objectives, such as average completion time, makespan, or cloud computing resource cost, by up to 3.5x. I will conclude the talk with discussion on future directions for optimizing Machine Learning systems.

BIO:

Deepak Narayanan is a final-year PhD student at Stanford University advised by Prof. Matei Zaharia. He is interested in designing and building software to improve the runtime performance and efficiency of emerging machine learning and data analytics workloads on modern hardware. His work is supported by a NSF graduate fellowship.

Yale University